Benefits of High Speed Interconnects to Cluster File Systems: A Case Study with Lustre

W. Yu, R. Noronha, S. Liang and D. K. Panda
Dept of Computer Sci. and Engg.
The Ohio State University
{yuw,noronha,liangs,panda}@cse.ohio-state.edu

Presented by Pavan Balaji

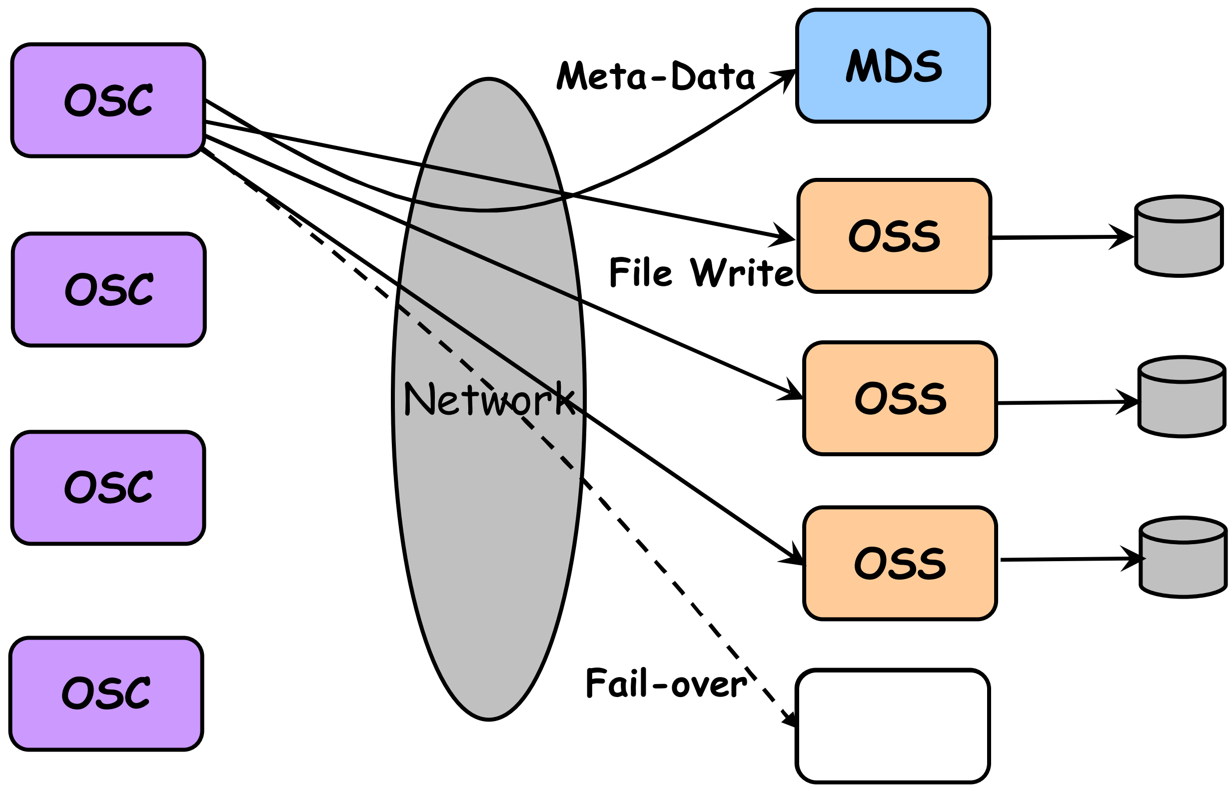
Data-Intensive Applications

- Recent trends in scientific applications
  - Compute → Data-intensive (tera-bytes to peta-bytes)
  - Disks are significantly slower than memory and networks
- Solution: Parallel file systems
  - E.g., Lustre, PVFS, Panasas, pNFS
  - General Idea: Perform I/O in parallel on multiple nodes
- Network capability: critical for performance
- Lustre file-system is of particular interest
  - Designed and developed by Cluster File Systems

Overview of Lustre

- Lustre Parallel File System
  - POSIX compliant, stateful, object-based file system
- Three Important Subsystems:
  - Object Storage Client (OSC)
  - Meta-data Server (MDS)
  - Object Storage Server (OSS)
- Support for Fault Tolerance
  - MDS fault tolerance available
  - OSS fault tolerance upcoming

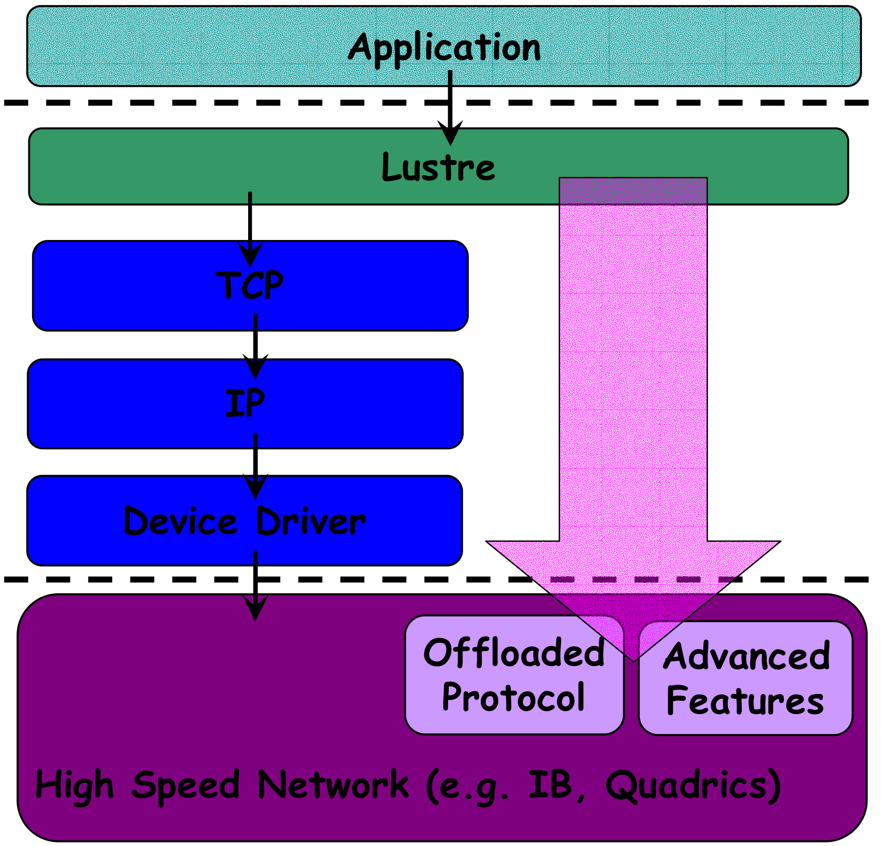
Lustre Architecture



InfiniBand and Quadrics

- InfiniBand
  - An emerging industry standard
  - Delivers low latency (2us) and high bandwidth (8Gbps)
  - Advanced Features: RDMA, Multicast, QoS, Atomic Operations
- Quadrics
  - High performance (10Gbps), low latency (<2us)
  - Advanced Features: QDMA, RDMA
  - Intelligent Switch Support
  - Programmable Network Adapter

Lustre over High Speed Interconnects



- Lustre over TCP/IP
  - Generic Solution
  - Sub-optimal Performance
- Lustre over native network
  - N/w specific implementation
  - Modified for IB or Quadrics
  - Performance Improvement?

Objectives

- Which file system operations of Lustre can benefit more from the capabilities of the native protocols on high speed interconnects?
- What are the aspects of Lustre that need to be further strengthened?
- Can the latest I/O-bus technologies, such as PCI-Express, help the performance of Lustre?

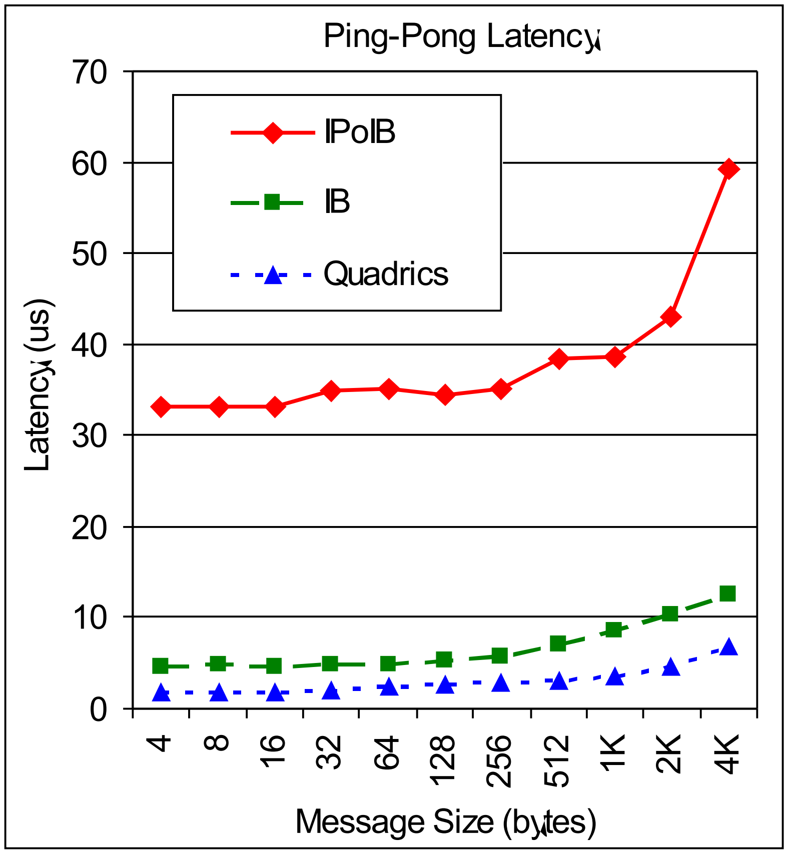
Presentation Outline

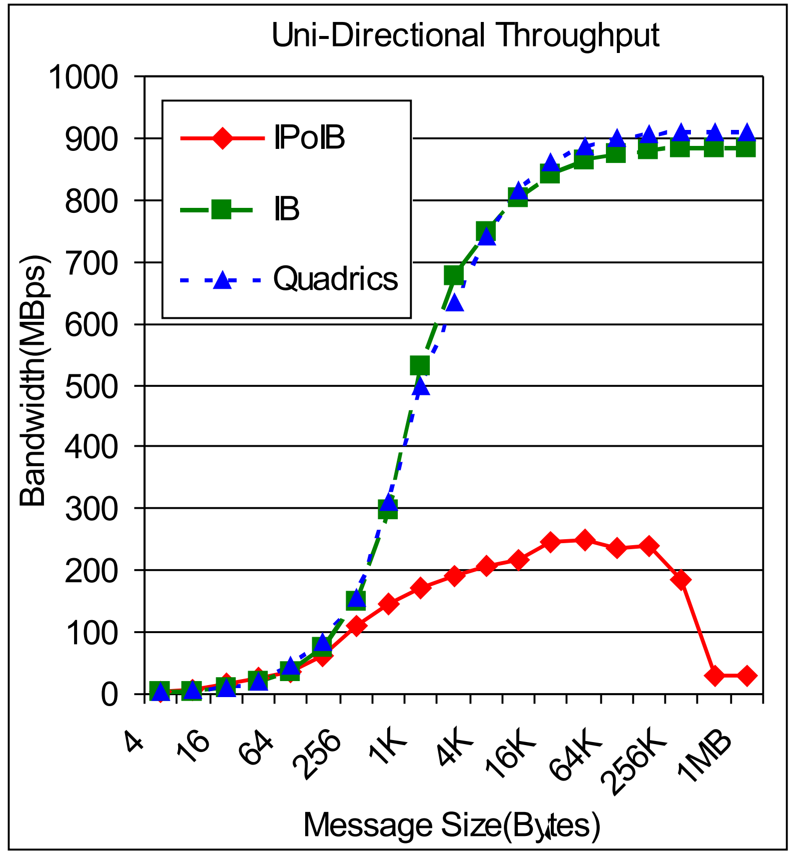
- Overview of Lustre and High Speed Interconnects (IBA and Quadrics)
- Performance Evaluation (IB, Quadrics, TCP/IP/IB)
  - Basic Network Performance
  - Sequential I/O Performance
  - Parallel I/O Performance
  - Benefits of PCI-Express
- Conclusions and Future Work

Experiment Testbeds

- Cluster 1: Eight-nodes
  - Dual Intel Xeon 3.0GHz; PCI-X 133Mhz/64bit; 1GB DDR, 512KB L2 cache
- Cluster 2: Four-nodes
  - Dual Intel EM64T 3.4Ghz; x8 PCI-Express and PCI-X 133Mhz/64bit
  - 1GB DDR, 1024KB L2 cache
- Networks:
  - IB: MT23108 (PCI-X) and MT25208 (PCI-Ex)
    - A 144-port InfiniScale switch
  - Quadrics: QS-8A switch, Elan4 cards

Basic Performance Comparison



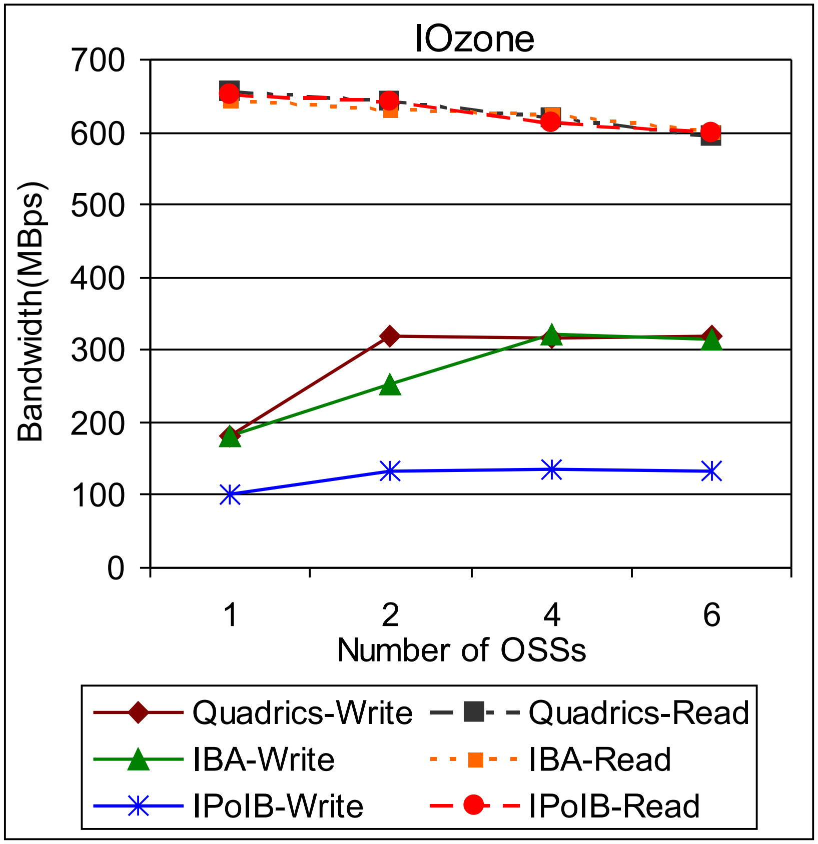


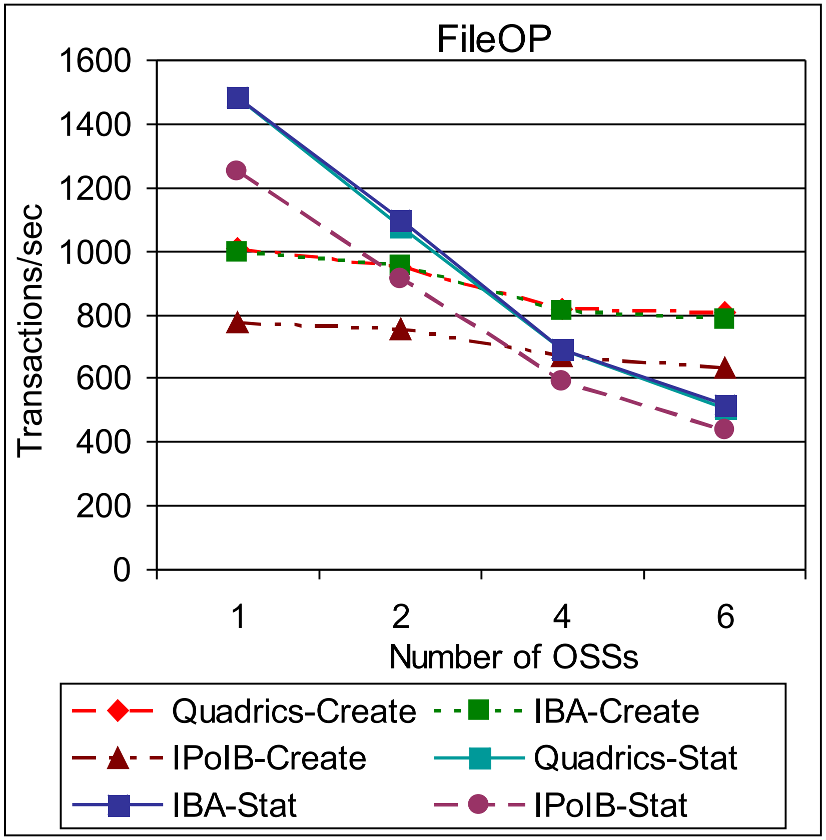
- 2 to 4 times improvement compared to IPoIB
- Comparable performance for native IB and native Quadrics

Presentation Outline

- Overview of Lustre and High Speed Interconnects (IBA and Quadrics)
- Performance Evaluation (IB, Quadrics, TCP/IP/IB)
  - Basic Network Performance
  - Sequential I/O Performance
  - Parallel I/O Performance
  - Benefits of PCI-Express
- Conclusions and Future Work

Read/Write and FileOP





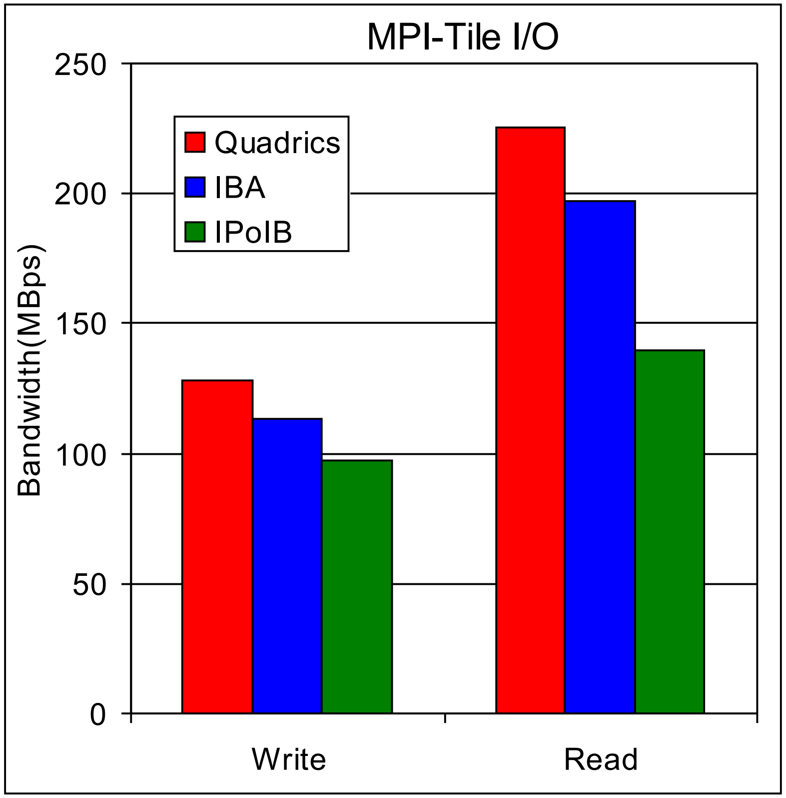
- Reads are largely cached for IOZone
- Some operations (especially FileOPs) do not scale very well with increasing OSSs

Postmark Application

Table 1. Postmark Performance (Trans/Sec)			
OSS	Quadrics	IBA	IPoIB
1	500	320	283
2	250	220	170
4	186	177	132
6	150	153	113

- Postmark involves mostly large volume of small file write
- Lower latency of Quadrics also help postmark performance

MPI-Tile I/O and BT/IO



BT/IO

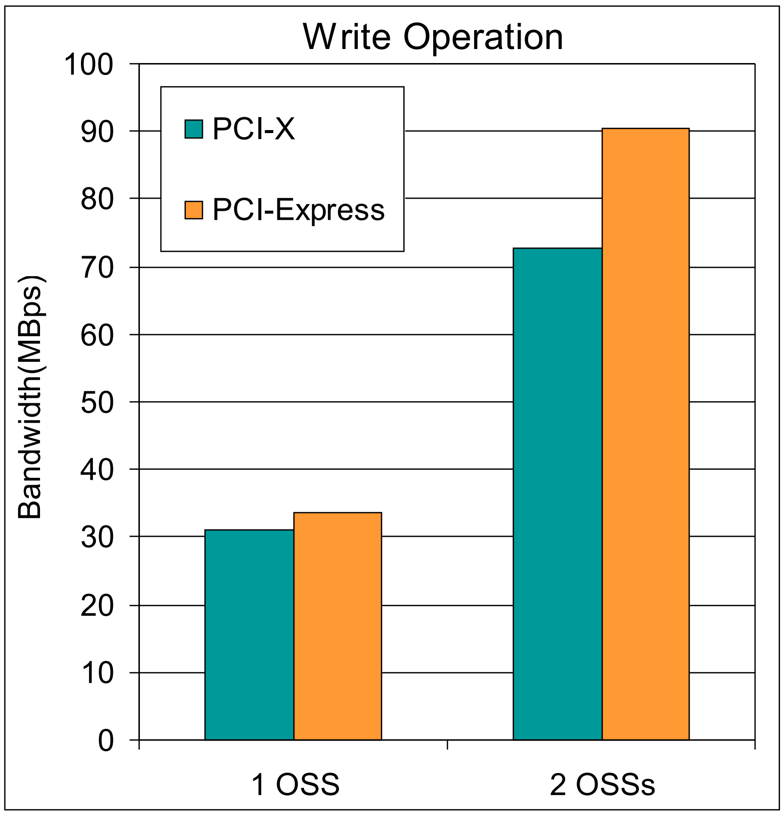
Type	Duration (sec)	IO Time (sec)
BT	61.34	--
BT/IO Quadrics	69.08	7.74
BT/IO IBA	69.11	7.77
BT/IO IPoIB	73.59	12.25

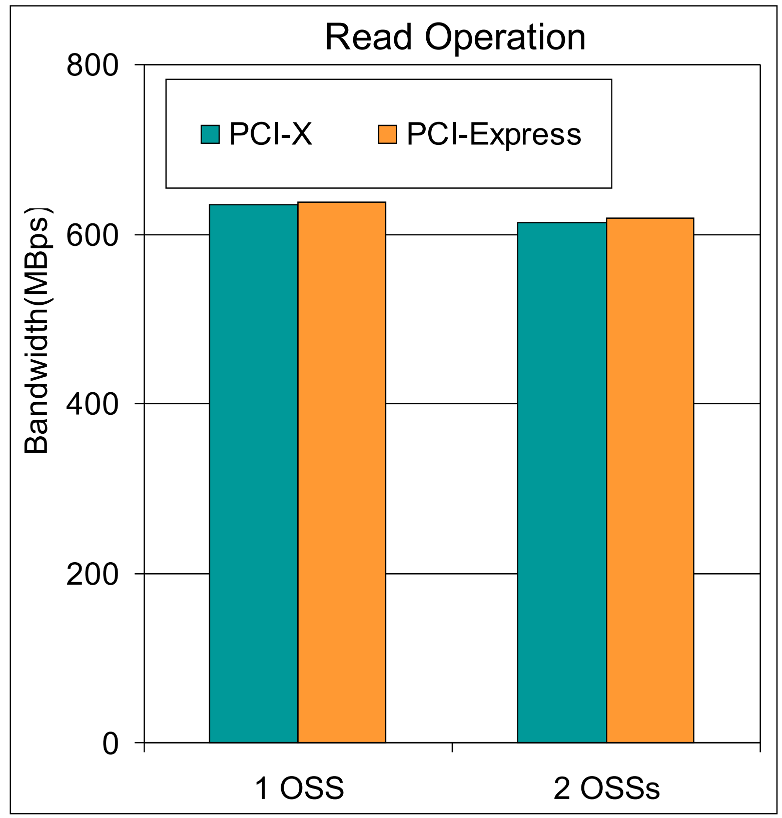
- Non-contiguous I/O in MPI-Tile-IO leads to small I/O operations
- Native implementations outperform IPoIB by twice for BT/IO

Presentation Outline

- Overview of Lustre and High Speed Interconnects (IBA and Quadrics)
- Performance Evaluation (IB, Quadrics, TCP/IP/IB)
  - Basic Network Performance
  - Sequential I/O Performance
  - Parallel I/O Performance
  - Benefits of PCI-Express
- Conclusions and Future Work

Benefits of PCI-Express





- Read requests are mostly cached
- Some improvement for Write Operations

Presentation Outline

- Overview of Lustre and High Speed Interconnects (IBA and Quadrics)
- Performance Evaluation (IB, Quadrics, TCP/IP/IB)
  - Basic Network Performance
  - Sequential I/O Performance
  - Parallel I/O Performance
  - Benefits of PCI-Express
- Conclusions and Future Work

Conclusions

- Compared the performance of Lustre over TCP/IP with native implementations over IB and Quadrics
  - Native implementations of IB and Quadrics perform about twice as better than over TCP/IP
  - Comparable performance results were observed between the native implementations over InfiniBand and Quadrics
- Scalability with increasing number of OSSs is not the best – further improvement is necessary
- InfiniBand blended with PCI-Express technology can further provide more performance advantages

Future Work

- Evaluate performance of Lustre over larger clusters
- Optimize Lustre with scalable Meta-data management – Meta Data Parallelization
- Evaluation with the Sockets Direct Protocol (SDP)
- Applications such as Checkpoint/Restart are I/O intensive – Design and Evaluation in such environments

Acknowledgements

- Current Funding support by

Office of Science U.S. DEPARTMENT OF ENERGY

NSF

CISCO SYSTEMS EMPOWERING THE INTERNET GENERATION

- Current Equipment support by

intel.

Mellanox Technologies

Web Pointers

NBCL

Group Webpage: http://nowlab.cse.ohio-state.edu

Project Page: http://nowlab.cse.ohio-state.edu/projects/clust-storage/

Emails: {yuw, noronha, liangs, panda}@cse.ohio-state.edu